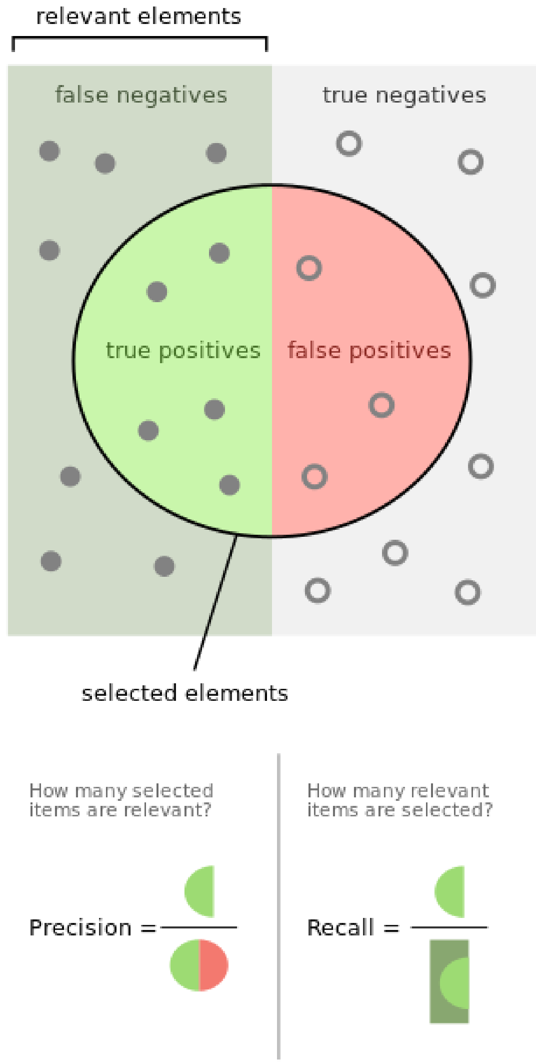
WIKIPEDIA

Precision and recall

In pattern recognition, information retrieval and binary classification, **precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while **recall** (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.



Precision and recall

Suppose a computer program for recognizing dogs in photographs identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 identified as dogs, 5 actually are dogs (true positives), while the rest are cats (false positives). The program's precision is 5/8 while its recall is 5/12. When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3. So, in this case, precision is "how useful the search results are", and recall is "how complete the results are".

In statistics, if the null hypothesis is that all items are *irrelevant* (where the hypothesis is accepted or rejected based on the number selected compared with the sample size), absence of type I and type II errors corresponds respectively to maximum precision (no false positive) and maximum recall (no false negative). The above pattern recognition example contained 8 − 5 = 3 type I errors and 12 − 5 = 7 type II errors. Precision can be seen as a measure of exactness or *quality*, whereas recall is a measure of completeness or *quantity*.

In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results.

Contents

Introduction

In an information retrieval scenario, the instances are documents and the task is to return a set of relevant documents given a search term; or equivalently, to assign each document to one of two categories, "relevant" and "not relevant". In this case, the "relevant" documents are simply those that belong to the "relevant" category. Recall is defined as the *number of relevant documents* retrieved by a search *divided by the total number of existing relevant documents*, while precision is defined as the *number of relevant documents* retrieved by a search *divided by the total number of documents retrieved* by that search.

In a classification task, the precision for a class is the *number of true positives* (i.e. the number of items correctly labeled as belonging to the positive class) *divided by the total number of elements labeled as belonging to the positive class* (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the *number of true positives divided by the total number of elements that actually belong to the positive class* (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

In information retrieval, a perfect precision score of 1.0 means that every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved) whereas a perfect recall score of 1.0 means that all relevant documents were retrieved by the search (but says nothing about how many irrelevant documents were also retrieved).

In a classification task, a precision score of 1.0 for a class C means that every item labeled as belonging to class C does indeed belong to class C (but says nothing about the number of items from class C that were not labeled correctly) whereas a recall of 1.0 means that every item from class C was labeled as belonging to class C (but says nothing about how many other items were incorrectly also labeled as belonging to class C). [which items]

Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. Brain surgery provides an illustrative example of the tradeoff. Consider a brain surgeon tasked with removing a cancerous tumor from a patient's brain. The surgeon needs to remove all of the tumor cells since any remaining cancer cells will regenerate the tumor. Conversely, the surgeon must not remove healthy brain cells since that would leave the patient with impaired brain function. The surgeon may be more liberal in the area of the brain she removes to ensure she has extracted all the cancer cells. This decision increases recall but reduces precision. On the other hand, the surgeon may be more conservative in the brain she removes to ensure she extracts only cancer cells. This decision increases precision but reduces recall. That is to say, greater recall increases the chances of removing healthy cells (negative outcome) and increases the chances of removing all cancer cells (positive outcome). Greater precision decreases the chances of removing healthy cells (positive outcome) but also decreases the chances of removing all cancer cells (negative outcome).

Usually, precision and recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. *precision at a recall level of 0.75*) or both are combined into a single measure. Examples of measures that are a combination of precision and recall are the F-measure (the weighted harmonic mean of precision and recall), or the Matthews correlation coefficient, which is a geometric mean of the chance-corrected variants: the regression coefficients Informedness (DeltaP') and Markedness (DeltaP).[1][2] Accuracy is a weighted arithmetic mean of Precision and Inverse Precision (weighted by Bias) as well as a weighted arithmetic mean of Recall and Inverse Recall (weighted by Prevalence).[1] Inverse Precision and Inverse Recall are simply the Precision and Recall of the inverse problem where positive and negative labels are exchanged (for both real classes and prediction labels). Recall and Inverse Recall, or equivalently true positive rate and false positive rate, are frequently plotted against each other as ROC curves and provide a principled mechanism to explore operating point tradeoffs. Outside of Information Retrieval, the application of Recall, Precision and F-measure are argued to be flawed as they ignore the true negative cell of the contingency table, and they are easily manipulated by biasing the predictions.[1] The first problem is 'solved' by using Accuracy and the second problem is 'solved' by discounting the chance component and renormalizing to Cohen's kappa, but this no longer affords the opportunity to explore tradeoffs graphically. However, Informedness and Markedness are Kappa-like renormalizations of Recall and Precision,[3] and their geometric mean Matthews correlation coefficient thus acts like a debiased F-measure.

Definition (information retrieval context)

In information retrieval contexts, precision and recall are defined in terms of a set of *retrieved documents* (e.g. the list of documents produced by a web search engine for a query) and a set of *relevant documents* (e.g. the list of all documents on the internet that are relevant for a certain topic), cf. relevance. The measures were defined in Perry, Kent & Berry (1955)

Precision

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

For example, for a text search on a set of documents, precision is the number of correct results divided by the number of all returned results.

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called *precision at n* or *P@n*.

Precision is used with recall, the percent of *all* relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.

Note that the meaning and usage of "precision" in the field of information retrieval differs from the definition of accuracy and precision within other branches of science and technology

Recall

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

For example, for a text search on a set of documents, recall is the number of correct results divided by the number of results that should have been returned.

In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the precision.

Definition (classification context)

For classification tasks, the terms *true positives*, *true negatives*, *false positives*, and *false negatives* (see Type I and type II errors for definitions) compare the results of the classifier under test with trusted external judgments. The terms *positive* and *negative* refer to the classifier's prediction (sometimes known as the *expectation*), and the terms *true* and *false* refer to whether that prediction corresponds to the external judgment (sometimes known as the *observation*).

Let us define an experiment from *P* positive instances and *N* negative instances for some condition. The four outcomes can be formulated in a 2×2 contingency table or confusion matrix, as follows:

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Precision and recall are then defined as:[6]

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Recall in this context is also referred to as the true positive rate or sensitivity, and precision is also referred to as positive predictive value (PPV); other related measures used in classification include true negative rate and accuracy.[6] True negative rate is also calledspecificity.

$$\text{True negative rate} = \frac{tn}{tn + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Additionally, the predicted positive condition rate (PPCR) identifies the percentage of the total population that is flagged; for example, for a search engine returning 30 results (retrieved documents) out of 1,000,000 documents, the PPCR is 0.003%.

$$\text{Predicted positive condition rate} = \frac{}{tp +}$$

Probabilistic interpretation

It is possible to interpret precision and recall not as ratios but as probabilities:

- Precision is the probability that a (randomly selected) retrieved document is relevant.
- Recall is the probability that a (randomly selected) relevant document is retrieved in a search.

Note that the random selection refers to a uniform distribution over the appropriate pool of documents; i.e. by *randomly selected retrieved document*, we mean selecting a document from the set of retrieved documents in a random fashion. The random selection should be such that all documents in the set are equally likely to be selected.

Note that, in a typical classification system, the probability that a retrieved document is relevant depends on the document. The above interpretation extends to that scenario also (needs explanation).

Another interpretation for precision and recall is as follows. Precision is the average probability of relevant retrieval. Recall is the average probability of complete retrieval. Here we average over multiple retrieval queries.

F-measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

Terminology and derivations from a confusion matrix

condition positive (P)
the number of real positive cases in the data
condition negative (N)
the number of real negative cases in the data

---

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

---

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$$

miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

false discovery rate (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} = 1 - \mathrm{PPV}$$

false omission rate (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TN}} = 1 - \mathrm{NPV}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

F1 score
is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Informedness or Bookmaker Informedness (BM)

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$

Markedness (MK)

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

Sources: Fawcett (2006), Powers (2011), and Ting (2011) [4] [1] [5]

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

This measure is approximately the average of the two when they are close, and is more generally the harmonic mean, which, for the case of two numbers, coincides with the square of the geometric mean divided by the arithmetic mean. There are several reasons that the F-score can be criticized in particular circumstances due to its bias as an evaluation metric.[1] This is also known as the F_1 measure, because recall and precision are evenly weighted.

It is a special case of the general F_β measure (for non-negative real values of β):

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\beta^2 \cdot \mathrm{precision} + \mathrm{recall}}$$

Two other commonly used F measures are the F_2 measure, which weights recall higher than precision, and the $F_{0.5}$ measure, which puts more emphasis on precision than recall.

The F-measure was derived by van Rijsbergen (1979) so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure $E_\alpha = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$, the second term being the weighted harmonic mean of precision and recall with weights $(\alpha, 1 - \alpha)$. Their relationship is $F_\beta = 1 - E_\alpha$ where $\alpha = \frac{1}{1 + \beta^2}$.

Limitations as goals

There are other parameters and strategies for performance metric of information retrieval system, such as the area under the ROC curve (AUC).[7]

For web document retrieval, if the user's objectives are not clear, the precision and recall can't be optimized. As summarized by Lopresti,[8]

> Browsing is a comfortable and powerful paradigm (the serendipity effect).
>
> - Search results don't have to be very good.
> - Recall? Not important (as long as you get at least some good hits).
> - Precision? Not important (as long as at least some of the hits on the first page you return are good).

See also

- Uncertainty coefficient, also called *proficiency*
- Sensitivity and specificity

References

1. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf) (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
2. Perruchet, P.; Peereman, R. (2004). "The exploitation of distributional information in syllable processing". *J. Neurolinguistics*. **17** (2–3): 97–119. doi:10.1016/s0911-6044(03)00059-9 (https://doi.org/10.1016%2Fs0911-6044%2803%2900059-9).
3. Powers, David M. W. (2012). "The Problem with Kappa". *Conference of the European Chapter of the Association for Computational Linguistics (EACL2012) Joint ROBUS-UNSUP Workshop*.
4. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).
5. Ting, Kai Ming (2011). *Encyclopedia of machine learning* (https://link.springer.com/referencework/10.1007%2F978-0-387-30164-8). Springer. ISBN 978-0-387-30164-8.
6. Olson, David L.; and Delen, Dursun (2008); *Advanced Data Mining Techniques*, Springer, 1st edition (February 1, 2008), page 138, ISBN 3-540-76916-1
7. Zygmunt Zając. What you wanted to know about AUC. http://fastml.com/what-you-wanted-to-know-about-auc/
8. Lopresti, Daniel (2001); *WDA 2001 panel* (http://www.csc.liv.ac.uk/~wda2001/Panel_Presentations/Lopresti/Lopresti_files/v3_document.htm)

- Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999). *Modern Information Retrieval*. New York, NY: ACM Press, Addison-Wesley, Seiten 75 ff. ISBN 0-201-39829-X

- Hjørland, Birger (2010); *The foundation of the concept of relevance*, Journal of the American Society for Information Science and Technology, 61(2), 217-237
- Makhoul, John; Kubala, Francis; Schwartz, Richard; and Weischedel, Ralph (1999); *Performance measures for information extraction*, in *Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999*
- Perry, James W.; Kent, Allen; Berry, Madeline M. (1955). "Machine literature searching X. Machine language; factors underlying its design and development". *American Documentation* **6** (4): 242. doi:10.1002/asi.5090060411.
- van Rijsbergen, Cornelis Joost "Keith" (1979); *Information Retrieval*, London, GB; Boston, MA: Butterworth, 2nd Edition, ISBN 0-408-70929-4

External links

- Information Retrieval – C. J. van Rijsbergen 1979
- Computing Precision and Recall for a Multi-class Classification Problem

Retrieved from "https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=853202943"

This page was last edited on 3 August 2018, at 03:32 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.